Constructing a data-driven receptor model for organic and inorganic aerosol : a synthesis analysis of eight mass spectrometric data sets from a boreal forest site

Äijälä, Mikko

2019-03-21

Äijälä , M , Dällenbach , K , Canonaco , F , Heikkinen , L , Junninen , H , Petäjä , T , Kulmala , M , Prevot , A S H & Ehn , M 2019 , ' Constructing a data-driven receptor model for organic and inorganic aerosol : a synthesis analysis of eight mass spectrometric data sets from a boreal forest site ' , Atmospheric Chemistry and Physics , vol. 19 , no. 6 , pp. 3645-3672 . https://doi.org/10.5194/acp

http://hdl.handle.net/10138/302087
https://doi.org/10.5194/acp-19-3645-2019

Atmos. Chem. Phys., 19, 3645–3672, 2019
https://doi.org/10.5194/acp-19-3645-2019

Constructing a data-driven receptor model for organic and inorganic aerosol – a synthesis analysis of eight mass spectrometric data sets from a boreal forest site

Mikko Äijälä [1], Kaspar R. Daellenbach[1], Francesco Canonaco[2], Liine Heikkinen [1], Heikki Junninen [1,3], Tuukka Petäjä[1], Markku Kulmala [1], André S. H. Prévôt[2], and Mikael Ehn[1]

[1]Institute for Atmospheric and Earth System Research/Physics, University of Helsinki, Helsinki, Finland
[2]Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, Villigen, Switzerland
[3]Laboratory of Environmental Physics, University of Tartu, Tartu, Estonia

Correspondence:Mikael Ehn (mikael.ehn@helsinki.)

Received: 6 October 2018 – Discussion started: 19 October 2018
Revised: 18 January 2019 – Accepted: 15 February 2019 – Published: 21 March 2019

Abstract. The interactions between organic and inorganic aerosol chemical components are integral to understanding and modelling climate and health-relevant aerosol physicochemical properties, such as volatility, hygroscopicity, light scattering and toxicity. This study presents a synthesis analysis for eight data sets, of non-refractory aerosol composition, measured at a boreal forest site. The measurements, performed with an aerosol mass spectrometer, cover in total around 9 months over the course of 3 years. In our statistical analysis, we use the complete organic and inorganic unit-resolution mass spectra, as opposed to the more common approach of only including the organic fraction. The analysis is based on iterative, combined use of (1) data reduction, (2) classi cation and (3) scaling tools, producing a data-driven chemical mass balance type of model capable of describing site-speci c aerosol composition. The receptor model we constructed was able to explain 83 % of variation in data, which increased to 96 3 % when signals from low signal-to-noise variables were not considered. The resulting interpretation of an extensive set of aerosol mass spectrometric data infers seven distinct aerosol chemical components for a rural boreal forest site: ammonium sulfate (35 7 % of mass), low and semi-volatile oxidised organic aerosols (27 8 % and 12 7 %), biomass burning organic aerosol (11 7 %), a nitrate-containing organic aerosol type (7 2 %), ammonium nitrate (5 2 %), and hydrocarbon-like organic aerosol (3 1 %). Some of the additionally observed, rare outlier aerosol types likely emerge due to surface ionisation effects and likely represent amine compounds from an unknown source and alkaline metals from emissions of a nearby district heating plant. Compared to traditional, ion-balance-based inorganics apportionment schemes for aerosol mass spectrometer data, our statistics-based method provides an improved, more robust approach, yielding readily useful information for the modelling of submicron atmospheric aerosols physical and chemical properties. The results also shed light on the division between organic and inorganic aerosol types and dynamics of salt formation in aerosol. Equally importantly, the combined methodology exemplifies an iterative analysis, using consequent analysis steps by a combination of statistical methods. Such an approach offers new ways to home in on physicochemically sensible solutions with minimal need for a priori information or analyst interference. We therefore suggest that similar statistics-based approaches offer signi cant potential for un- or semi-supervised machine-learning applications in future analyses of aerosol mass spectrometric data.

1 Introduction

Along with particle size, aerosol chemical composition is fundamental in understanding aerosol physicochemical properties such as hygroscopicity, volatility, optics and toxicity (Bilde et al., 2015; Swietlicki et al., 2008; Zimmermann, 2015). In the past decade aerosol mass spectrometry has provided a way to quantitatively resolve basic chemical com-

position of aerosol in near real time. This not only enables basic chemical speciation into organic and common inorganic ion species, but also produces a wealth of complex mass spectrometric data. It has since become clear that these data sets, although super cially hard to interpret, are rich in chemical information and their statistical analysis yields considerable new knowledge. However, tapping into this information source requires use of advanced analysis tools and chemometric methods (i.e. "using mathematical and statistical methods to provide maximum chemical information by analysing chemical data"; Kowalski, 1975). Consequently, advanced statistical methods for data reduction have quickly gained traction in aerosol mass spectrometry, and are presently widely used for deconvolution of complex organic mass spectra into their underlying components. Speci cally, the positive matrix factorisation algorithm (PMF; Paatero and Tapper, 1994) has achieved a predominant status as the state-of-the-art analysis tool for deconvolving aerosol mass spectrometric data. Factorisation methods such as PMF notably allow for the condensation of information found in high-dimension data matrices into a manageable number of factors, corresponding to aerosol chemical species, sources or processes, for example. Data reduction often additionally allows for improved visualisation, aiding in interpretation of the underlying aerosol chemical phenomena.

In exploratory factor analysis, the principal dif culties often relate to deciding the optimal number of factors, choosing between multiple solutions of mathematically similar quality, and estimating the reliability and uncertainty of the results. Lacking robust but easy-to-use mathematical tools, the selection and interpretation of factorisation solutions remains prone to subjective bias by the analyst. Speci cally, while analyst-imposed additional constraints in factorisation may sometimes be required to reduce rotational uncertainty and extract minor factors in data (e.g. Canonaco et al., 2013; Crippa et al., 2014) such procedures are especially prone to analyst-subjective decisions. Evaluation and veri cation of a factorisation solution thus generally requires meticulous study and understanding of, for example, correlations with auxiliary data, temporal changes and cycles and spectral references. While statistics-driven methods for spectra comparison and classi cation as of yet remain marginal in aerosol mass spectrometry, they do show promise in their capability to automatically group similar spectra based on their chemically relevant features, producing comparable classi cations to those performed manually by expert analysts (Äijälä et al., 2017; Rebotier and Prather, 2007; Freutel et al., 2013).

The overwhelming majority of PMF analyses to date from aerosol mass spectrometer (AMS) have been performed on the organic fraction alone (Zhang et al., 2011). Contrary to popular belief, there exists no tenable reasons to limit chemometric analysis to organic signals, as exempli ed by the analyses of Sun et al. (2012) and Hao et al. (2014). Although it requires some additional data preparation and processing, inclusion of inorganics provides additional insight into, for example, salt formation in aerosol. In this work, we apply data reduction and classi cation methods for analysing organic and inorganic aerosol mass spectral data from several measurement campaigns in the boreal forest. We then derive a comprehensive receptor model resolving the dominant aerosol categories at the site. In addition, by presenting an example of a semi-supervised, statistics-driven analysis of large mass spectral data sets, we hope to pave the way for machine-learning-based data analysis approaches, reducing the need for expert analyst input and subjective judgement at each step.

2 Methods

Our instrumentation, data processing, measurement site and analysis algorithms have been comprehensively described in previous literature, to which we refer in the corresponding sections. Thus, we focus on the new aspects of this work, showing how the individual methods can be connected to form an analysis chain, and to exemplify how chemometric information can be propagated through it. In short, we will rst cover the measurement site, SMEAR II (Station for Measuring Ecosystem–Atmosphere Relations) and the sets of data available to us (Sect. 2.1). We then describe our mass spectrometer instrument and preparation of data (Sect. 2.2). In Sect. 2.3, we will brie y go through the various statistical tools and algorithms, covering the basics of data factorisation, classi cation of spectra using a clustering algorithm and clustering solution evaluation, and detail the pre- and post-weighting involved. Section 2.4 describes typical reference methods for inorganics and nitrate apportionment: an ion balance scheme and a separate parameterisation for estimating organonitrate loading, to provide a comparison for the inorganic speciation from our statistics-based receptor model. Finally, in Sect. 2.5, we present a summarised, step-by-step description of how the methods were combined to produce a receptor model for aerosol composition at the measurement site.

2.1 Measurement site and collection of data

2.1.1 The SMEAR II site

The AMS data of this study were collected at the SMEAR II site in Hyytiälä, southern Finland (61°51′00″ N, 24°17′013″ E). The site is a well-known and well-equipped atmospheric research station, representing rural, background atmosphere in the boreal forest biome. The site and earlier measurements therein have been extensively described and reported in the literature (e.g. Hari and Kulmala, 2005; Williams et al., 2011; Äijälä et al., 2017).

The environment consists mostly of forests dominated by Scots pine (*Pinus Sylvestris*) – 90 % of land in the nearest 50 km, and 94 % in the nearest 5 km is forested (Williams et al., 2011).

A large part of the aerosol loading at SMEAR II is attributable to regional biogenic secondary organic aerosol (SOA; Corrigan et al., 2013; Crippa et al., 2014; Allan et al., 2006) and long-range transport from industrial regions in southern Finland, western Russia and central Europe (Kulmala et al., 2000; Patokoski et al., 2015; Niemi et al., 2009; Sogacheva et al., 2005). Regional anthropogenic aerosol sources include the towns Orivesi (pop. 9500; 19 km south) and Tampere (pop. 213 000; 48 km south-west), as well as two sawmills and a pellet factory in the village of Korkeakoski, Juupajoki (7 km east-south-east of the station). The surrounding countryside is sparsely populated (5–10 inhabitants km^2), and although emissions from agriculture, traf c, domestic heating, cooking and other combustion sources (saunas, barbecues, agricultural machinery etc.) are limited, they are clearly observable at the station and may increase aerosol loading in often plume-type pollution events. The anthropogenic organic aerosols were further analysed previously (Äijälä et al., 2017).

2.1.2 Data sets

In this study, the aerosol composition was monitored by an AMS between 2008 and 2011, during several short measurement campaigns. Notable larger, intensive campaigns at the time were the EUCAARI project (2008–2009; Kulmala et al., 2009, 2011) and HUMPPA-COPEC (2010; Williams et al., 2011; Corrigan et al., 2013). The sets of data used along with their time frames are shown in Table 1. Data availability by year and month is presented in Table 2.

2.2 Instrumentation, data processing and preparation

2.2.1 The aerosol mass spectrometer (AMS) instrument and basic data processing

The mass spectrometric data for this study were acquired with a Time-of-Flight Aerosol Mass Spectrometer (ToF-AMS), developed by Aerodyne Research Inc. (Billerica, MA, US). AMS instruments in general have been described by Canagaratna et al. (2007), and the compact ToF analyser version (CToF) used in this study by Drewnick et al. (2005). Additional, more speci c details related to the speci c instrument we used are available in our previous study (Äijälä et al., 2017).

In brief, the AMS instrument sucks sample aerosol from atmospheric pressure to vacuum conditions through an inlet system consisting of a critical ori ce and a particle concentrating aerodynamic lens (Liu et al., 2007). The sample aerosol beam is directed at a vaporiser operated at 600 °C, whereby ash vaporisation of non-refractory aerosol components occurs. The resulting vapour is ionised using 70 eV electron impact ionisation – a well-characterised hard ionisation technique, resulting in rather universal and predictable, but highly fragmenting ionisation. Finally, the ions are led to an orthogonal extraction re ectron time-of- ight mass analyser, where the ions' mass-to-charge (m/z) ratios are measured.

The per-amu (atomic mass unit) analyser signal is subsequently quanti ed based on instrument response calibrations and corrections (among others the correction for relative ionisation ef ciency between the species, RIE; Allan et al., 2004; Supplement Sect. S4). Individual, unit-mass-resolution amu signals are then chemically speciated, based on chemical information on fragmentation and air composition (see Allan et al., 2003b, for details). Additional, speci c minor modi cations to our instrument have been discussed in our previous work (Äijälä et al., 2017).

2.2.2 Data preparation and down-weighting

After basic processing, the data were further prepared, to serve as input for factorisation (described in following Sect. 2.3). The organic and inorganic data and related uncertainties were extracted, and down-weighting of signals performed. The procedure for extraction and preparation of AMS organic signal and related error matrices has been described by Allan et al. (2003b) and Ulbrich et al. (2009). In short, measurement points or variables with missing data were omitted and error matrices calculated, based on a function accounting for both counting-statistics-induced uncertainty as well as background noise from the detector and electronics. The signals were then down-weighted by multiplying the error-matrix-conveyed uncertainty values for low signal-to-noise ratio (SNR) variables with a scalar: "weak" variables (SNR < 3) were down-weighted by a factor of 2 and "bad" variables (SNR < 1) by 10. The procedure for inorganics (SO_4, NO_3, NH_4, Chl, i.e. sulfates, nitrates, ammonia and chloride species) was similar to that used for the organics ("org"), including for the down-weighting of signals derived from fragmentation calculations. Analogous to the basic procedure of down-weighting "duplicate information" organic signals, e.g. those derived from m/z 44 Th (mainly CO_2^+), similarly derived inorganic signal weights were normalised so that their weight of the original plus "duplicate" signals equalled that of the original signal. Finally, the matrices for all the ion species (org, SO_4, NO_3, NH_4, Chl, in nitrate equivalent mass) were combined to form the nal input matrices for factorisation, while retaining speciation information in the ion indexing.

2.3 Statistical methods and algorithms

2.3.1 Positive matrix factorisation

For factorisation, we used the PMF model developed by Pentti Paatero and colleagues (Paatero, 1997, 1999; Paatero and Tapper, 1994) and widely used for analysis of AMS data since 2007 (Lanz et al., 2007b; Zhang et al., 2011). In brief, PMF is a statistical model, typically resolving a bilinear lin-

Table 1. Data sets used in this study and their time frames (dd.mm.yyyy).

Data set number	Data set name	Campaign	Start time	End time
I	"May 2008"	EUCAARI	29.04.2008	08.06.2008
II	"Sep 2008"	EUCAARI	10.09.2008	15.10.2008
III	"Mar 2009"	EUCAARI	04.03.2009	06.04.2009
IV	"May 2009"		29.04.2009	28.05.2009
V	"Jun 2009"		12.06.2009	08.08.2009
VI	"Aug 2009"		13.08.2009	19.09.2009
VII	"Summer 2010"	HUMPPA-COPEC	09.07.2010	07.08.2010
VIII	"Winter 2010"		10.11.2010	07.01.2011

Table 2. For months when AMS data were available, percentages indicate the fraction of days with at least one data point.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2008	–	–	–	–	65 %	20 %	–	–	70 %	48 %	–	–
2009	–	–	94 %	23 %	90 %	63 %	81 %	87 %	63 %	–	–	–
2010	–	–	–	–	–	–	74 %	68 %	–	–	47 %	100 %
2011	23 %	–	–	–	–	–	–	–	–	–	–	–

ear combination of factor pro les (G) and time series (F) best describing the measured data matrix (X, Eq. 1). The residual matrix E then denotes the portion of data left unexplained by the model (i.e. residual). The PMF model is thus formulated:

$$X_{(t,v)} = G_{(t,f)} F_{(f,v)} + E_{(t,v)}. \quad (1)$$

The brackets indicate matrix dimensions, with v denoting number of variables, t the number of time points and f the number of factors. As shown in Eq. (1), the model can be solved for any f ($< v, t$), requiring it to be selected by the analyst.

The main features setting PMF apart from other similar factorisation models, and making it particularly suitable for atmospheric aerosol models, are on the one hand the limitation of factor pro les and time series to positive values, hence drastically reducing the amount of rotational ambiguity, and on the other hand the improved error model where the quantity to minimise is the weighted (typically the measurement uncertainty) residual, resulting in higher weight for the variables with better SNR. In PMF, the minimum weighted residual is solved using one of the related algorithms, i.e. PMF2 or Multilinear Engine 2 (ME-2; Paatero, 1999). Of the two algorithms, ME-2 can take in additional equations de ned by the user, i.e. constraints the solutions need to adhere to. In this study, when ME-2 constraints were applied to the factor pro les, we set upper and lower bounds for the allowed pro le solutions. The bounds were based on variability estimates obtained from earlier analysis, as explained later, in Sect. 2.5. Variability estimate of the nal model is available in the Supplement (Fig. S13). For running the PMF and ME-2 algorithms, we used the Igor Pro (Wavemetrics Inc.) based SoFi (v. 4.8) user interface developed by Francesco Canonaco and co-workers at Paul Scherrer Institute (PSI). The interface allows input of the pre-processed data and user-selected parameters, and calls on the solver algorithms (PMF2 or ME-2, depending on assignment) to return a solution to be displayed and analysed in SoFi (Canonaco et al., 2013).

When PMF is used as a standalone method for source attribution, the selection of solution needs to be carefully validated. Sensitivities towards a different number of factors, rotations and initialisation seeds are meticulously analysed, and correlations with auxiliary data are computed. A case is then made for why the selection is the best possible. Contrarily, in our analysis approach, we do not claim to arrive at optimum solutions for individual PMF–ME-2 runs. Instead, we rely on a multitude of data de-convolution runs to uncover the main structures in the ensemble of all data sets, and use statistical classi cation methods to evaluate the general outlook and commonalities between PMF–ME-2 factors at each analysis phase. As discussed in Sect. 2.5, this trade-off instead enables us to concentrate on best modelling the entirety of all data sets

2.3.2 Relaxed chemical mass balance model

To harmonise the description of aerosol components, we constructed a constrained receptor model, where all the pro le components were constrained. For this purpose we applied a ME-2-based chemical mass balance (CMB) type of model. CMB models are typically used as receptor models for cases where source pro les are known, and only the mass loading information needs resolving (Friedlander, 1973; Gordon, 1988; Hopke, 1991, 2016; Miller et al., 1972). In such mass-conservation-based models, the observed loadings are mod-

elled as a sum of multiple individual sources. Although CMB is often presented mathematically as the sum of loadings (Supplement; Sect. S1, Eq. S1), it can also be thought of as a special case of the bi-linear model described in Eq. (1). Only now the pro le matrix (F) is assumed xed, simplifying the problem to resolving the loading matrix (G) which minimises the residuals (E). CMB can be run using the SoFi interface, using the same ME-2 solver as for PMF and ME-2 applications (Canonaco et al., 2013).

In this work, we use a relaxed CMB-like bilinear model (henceforth abbreviated as r-CMB), where all the source pro les are constrained but allowed to vary within narrow limits (derived from variability estimates; see Sect. 2.5; Supplement Fig. S13). In strict technical terms this approach could be labelled "an extremely constrained ME-2 model", but we choose to use the term "relaxed CMB" to differentiate between the typical use of ME-2 or constraining only part of the pro les, which allows the model considerably more freedom. We regard our use of the model as much closer to the idea of constraining all pro les than (semi-)exploratory factorisation typical for ME-2. The naming also serves to better highlight the conceptual differences between models in the different analysis phases.

Generally, the biggest problems of the CMB models relate to the selection of source pro les, typically from spectral libraries, and handling of their uncertainty. In our use, the anchor spectra as well as the limits for their allowed variabilities are experimentally derived from data, alleviating some of these typical concerns.

2.3.3 k-means clustering

For spectra classi cation, we selected the k-means algorithm, speci cally because in our previous tests it was successful in classifying similar spectral data. The earlier tests additionally yielded useful information on the selection of the dissimilarity metric, as well as algorithm initialisation types and data weighting (Äijälä et al., 2017). The k-means algorithm (e.g. Ball and Hall, 1965; MacQueen, 1967; Steinhaus, 1956; Jain, 2010) is a rather simple, iterative algorithm that partitions a group of objects to a predesignated number of groups or "clusters" based on their relative distances (i.e. dissimilarities). For each iteration, the algorithm assigns all objects to their closest centroids, which are then re-calculated from the mean variable values of the objects in the updated clusters. The aim is to minimise the within-cluster sum of distance (variance) (J) between the objects' (C_n) locations (x_i) and the cluster centroid $_n$ they are assigned to (Eq. 2):

$$J(C_n) = \sum_{x_i \in C_n} \|x_i - {}_n\|^2. \quad (2)$$

The k-means algorithm iteratively converges on (any) minimum of total J (C) obtained by summing over all objects C_n. To increase chances of nding a global minimum, repetitions using different initialisations are used. Speci cally, we used the improved stepwise initialisation "kmeans++" (Arthur and Vassilvitskii, 2007; available in MATLAB v. 2017a for example, Math Works Inc., Natick, MA, USA).

2.3.4 Spectral similarity and mass scaling

Based on our earlier metric comparison (Äijälä et al., 2017), we used (Pearson) correlation as a metric for spectral dissimilarity (or "distance", d; Fortier and Solomon, 1966; Mcquitty, 1966):

$$d(u,v) = 1 - \frac{\sum_{i=1}^{n}(u_i - \bar{u})\cdot(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}\cdot\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}}; \quad (3)$$

where u and v are the spectra in vector form, with m/z variables as vector components, and $\bar{u}$ and $\bar{v}$ are the arithmetic mean values of u and v.

In clustering mass spectra, data weighting is often applied. Based on previous tests (Äijälä et al., 2017), we applied mass scaling of variables, advocated by Stein and Scott and others (Stein and Scott, 1994; Kim et al., 2012; Horai et al., 2010), giving additional emphasis to higher mass signals. This common practice is based on the idea that higher mass fragment ions are more indicative of their parent ions, and thus the original characteristic composition, while smaller fragments can be produced from a wider variety of molecular fragmentation events. In mass scaling the weighted variables (Q) are calculated by multiplying the original variables (x) by mass-to-charge-speci c weights (w), as presented in Eq. (4).

$$Q_{m/z} = x_{m/z} \cdot w_{m/z} = x_{m/z} \cdot m/z^{s_m}; \quad (4)$$

where the scaling factor s_m was optimised for each classi cation separately (Supplement; Sect. S2).

2.3.5 Silhouette metric and post-weighting

The optimisation of mass scaling was based on the silhouette metric (later also abbreviated as "silh"; Rousseeuw, 1987), ranging between 1 to 1 and providing a straightforward, quantitative way to evaluate performance of the classi cation algorithm. The object-speci c silhouette value (s_i) is de ned as

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i} & \text{for } a_i < b_i \\ 0 & \text{for } a_i = b_i \\ \frac{b_i}{a_i} - 1 & \text{for } a_i > b_i \end{cases}; \quad (5)$$

where a_i corresponds to the mean distance to other objects in the same cluster, and b_i similarly to the mean distance to objects in the nearest neighbouring cluster. A silhouette value close to unity indicates the object is well clustered, while a value close to zero indicates the classi cation is uncertain, and the point is likely situated in-between two possible centroids. A negative cluster value is indicatory of possible misclassi cation. Silhouette values can be calculated for any sin-

gle cluster as the arithmetic mean of the cluster members' silhouettes, or similarly as a mean over all objects, to evaluate the quality of the clustering solution as a whole.

In order to mitigate the k-means algorithm's known sensitivity to outliers, and to improve handling of between-cluster samples, we applied a simple post-processing to all cluster centroids and variability calculations: the centroid spectra and variabilities were calculated as weighted averages ($\bar{v}$), and weighted standard deviations (σ^2; Eq. 6) respectively, instead of the normal unweighted values (similar to Äijälä et al., 2017). As weights, we used the object specific silhouette values $s_i > 0$ (Eq. 5):

$$\bar{v} = \frac{\sum_{i=1}^{N} s_i v_i}{\sum_{i=1}^{N} s_i} \quad \sigma^2 = \frac{\sum_{i=1}^{N} s_i (v - \bar{v})^2}{\sum_{i=1}^{N} s_i}$$

$$s_i = \max(s_i, 0); \tag{6}$$

where v_i are the cluster member objects (spectra) This procedure down-weights likely misclassified objects (silhouette < 0) to zero, and penalises the more uncertain or questionable assignations (low silhouette) compared to the decidedly well-clustered objects (silhouette close to unity). Singleton clusters were omitted from this calculation, and their variability was thus left undefined.

Figure 1. Schematic representation of the inorganic apportionment scheme. The scheme is divided into three cases according to the ratio of $[NH_4^+]$ to $[SO_4^{2-}]$. $[NH_4^+]$ first combines with $[SO_4^{2-}]$ to form NH_4HSO_4 (Case 1), then further to $(NH_4)_2SO_4$ (Case 2). In these cases, any nitrate observed is considered organic. In Case 3 leftover $[NH_4^+]$

2.4 Standard approximations for aerosol inorganic speciation and organonitrate

2.4.1 Ion balance model for inorganics

Aerosol inorganic chemical speciation is better understood than the organic speciation, due to much lower diversity of the chemical compounds involved. A variety of aerosol inorganic equilibrium models exist and are typically used as modules in atmospheric meteorological and air quality models. However, performing thermodynamic equilibrium calculations is computationally demanding (e.g. Fountoukis and Nenes, 2007) and requires a good deal of auxiliary data on thermodynamic conditions and chemical activities. Due to the complexity of the models and increased data needs, simpler approximations are often used in connection with AMS inorganic speciation. In the following ion-balance-scheme description, we denote the respective AMS ion species molar concentrations in square brackets (e.g. $[NH_4^+]$, $[NO_3^-]$, $[SO_4^{2-}]$).

A typical salt formation approximation used for AMS results is the Hong et al. (2017) ion pairing scheme, used in aerosol volatility and light scattering models, for example (Hong et al., 2017; Zieger et al., 2015). The Hong et al. (2017) scheme is based on similar approximation of Gysel et al. (2007), which in turn is a simplification of the more extensive model by Reilly and Wood (1969). We modified the Hong et al. (2017) scheme to additionally allow organonitrate ($orgNO_3$) and speciate any leftover $[NH_4^+]$ as its own class